

Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets

Year: 2023 | Citations: 1814 | Authors: A. Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian

Abstract

We present Stable Video Diffusion - a latent video diffusion model for high-resolution, state-of-the-art text-to-video and image-to-video generation. Recently, latent diffusion models trained for 2D image synthesis have been turned into generative video models by inserting temporal layers and finetuning them on small, high-quality video datasets. However, training methods in the literature vary widely, and the field has yet to agree on a unified strategy for curating video data. In this paper, we identify and evaluate three different stages for successful training of video LDMs: text-to-image pretraining, video pretraining, and high-quality video finetuning. Furthermore, we demonstrate the necessity of a well-curated pretraining dataset for generating high-quality videos and present a systematic curation process to train a strong base model, including captioning and filtering strategies. We then explore the impact of finetuning our base model on high-quality data and train a text-to-video model that is competitive with closed-source video generation. We also show that our base model provides a powerful motion representation for downstream tasks such as image-to-video generation and adaptability to camera motion-specific LoRA modules. Finally, we demonstrate that our model provides a strong multi-view 3D-prior and can serve as a base to finetune a multi-view diffusion model that jointly generates multiple views of objects in a feedforward fashion, outperforming image-based methods at a fraction of their compute budget. We release code and model weights at <https://github.com/Stability-AI/generative-models>.